

Please answer each of the following problems. Refer to the course webpage for the **collaboration policy**, as well as for **helpful advice** for how to write up your solutions.

1. Probability refresher (4 points)

- (a) (1 point) What is the cardinality of the set of all subsets of $\{1, 2, \dots, n\}$? **[We are expecting a mathematical expression along with one or two sentences explaining why it is correct.]**

SOLUTION: There are two choices for each element: either it can be in the subset or not in the subset. Therefore, the number of possible subset choices is $2 \cdot 2 \cdots 2 = 2^n$.

- (b) (1 point) Suppose we choose a subset of $\{1, \dots, n\}$ uniformly at random: that is, every set has an equal probability of being chosen. Let X be a random variable denoting the cardinality (that is, the size) of a set randomly chosen in this way. Calculate the expected value of X and show your work. **[We are expecting a mathematically rigorous argument establishing your answer. Your solution should not include summation signs.]**

SOLUTION: We give three solutions here. Either is acceptable.

Let $A \subseteq \{1, \dots, n\}$ be a random subset. Notice that the size of A is distributed identically to the size of $\bar{A}$ (that is, the complement of A). So we have $E[|A|] = E[|\bar{A}|]$. But then we have

$$E[X] = E[|A|] = \frac{E[|A|] + E[|\bar{A}|]}{2} = \frac{E[|A| + |\bar{A}|]}{2} = \frac{n}{2},$$

using linearity of expectation. So the answer is $n/2$.

Here's a second way, which is a bit more straightforward and also relies heavily on linearity of expectation: Let x_i be an indicator variable which is 1 if i is in the random subset and 0 otherwise. Then $X = \sum_{i=1}^n x_i$, so by linearity of expectation $E[X] = E[\sum_{i=1}^n x_i] = \sum_{i=1}^n E[x_i]$. But because our subset is random, each element is included with probability $1/2$ and excluded with probability $1/2$, so $E[x_i] = \frac{1}{2}$ for all i . Then we can see that $E[X] = \sum_{i=1}^n \frac{1}{2} = \frac{n}{2}$.

Here's another way, which more directly uses the definition of the expectation of a variable: The expected value of X by definition is $E(X) = \sum_{U \in P(S)} \frac{1}{|P(S)|} |U|$ where $S = 1, 2, \dots, n$ and $P(S)$ is the power set of S . Let $z_{i,U}$ be an indicator variable where $z_{i,U} = 1$ iff $i \in U$. We can then rewrite $E(X)$ as:

$$E(X) = \sum_{U \in P(S)} \frac{1}{|P(S)|} \sum_{i \in S} z_{i,U}$$

Or equivalently,

$$E(X) = \sum_{i \in S} \frac{1}{|P(S)|} \sum_{U \in P(S)} z_{i,U}$$

Since any element appears exactly in half the number of all possible subsets of S , we know $\sum_{U \in P(S)} z_{i,U} = \frac{|P(S)|}{2}$ and get the following:

$$E(X) = \sum_{i \in S} \frac{1}{|P(S)|} \frac{|P(S)|}{2} = \sum_{i \in S} \frac{1}{2} = \frac{n}{2}$$

Therefore, the expected value of X is $\frac{n}{2}$.

- (c) (2 points) Let $rand(a, b)$ return an integer uniformly at random from the range $[a, b]$. Each call to $rand(a, b)$ is independent. Consider the following function:

```
f(k, n):
    if k <= 1:
        return rand(1, n)
    return 2 * f(k/2, n)
```

What is the expected value and variance of $f(k, n)$? Assume that k is a power of 2. Please show your work. **[In addition to your answer, we are expecting a mathematical derivation along with a brief (1-2 sentence) analysis of what the pseudocode above does in order to explain why your derivation is the right thing to do.]**

SOLUTION: First, we describe what the function $f(k, n)$ returns. We have

$$f(k, n) = 2 \cdot f(k/2, n) = 2^2 \cdot (f(k/4, n)) = \dots = 2^{\log_2(k)} (f(1, n)) = k \cdot f(1, n).$$

Thus, $f(k, n) = k \cdot X$, where X is a uniformly random integer in $[1, n]$. (That is, X is the output of $rand(1, n)$). Then

$$E[f(k, n)] = k \cdot E[X] = k \cdot \left(\frac{n+1}{2} \right),$$

where we have used that the expected value of $rand(1, n)$ is $\frac{n+1}{2}$.

Next we calculate the variance. By definition, $Var(X) = E[X^2] - E[X]^2$, so we have

$$Var(X) = \sum_{i=1}^n \frac{1}{n} i^2 - \left(\sum_{i=1}^n \frac{1}{n} i \right)^2$$

Using identities $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$, we can simplify

$$Var(X) = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2} \right)^2 = \frac{n^2 - 1}{12}$$

Finally,

$$\text{Var}(f(k, n)) = \text{Var}(k \cdot X) = k^2 \cdot \text{Var}(X) = \frac{k^2(n^2 - 1)}{12}.$$

2. **Fun with Big-O notation.** (6 points; 1 point each) Mark the following as **True** or **False**. Briefly but convincingly justify all of your answers, using the definitions of $O(\cdot)$, $\Theta(\cdot)$ and $\Omega(\cdot)$. [To see the level of detail we are expecting, the first question has been worked out for you.]

- (z) $n = \Omega(n^2)$. This statement is **False**. To see this, we will use a proof by contradiction. Suppose that, as per the definition of $\Omega(\cdot)$, there is some n_0 and some $c > 0$ so that for all $n \geq n_0$, $n \geq c \cdot n^2$. Choose $n = \max\{1/c, n_0\} + 1$. Then $n \geq n_0$, but we have $n > 1/c$, which implies that $c \cdot n^2 > n$. This is a contradiction.
- (a) $n = O(n \log(n))$.

SOLUTION: True. Consider the values $c = 1, n_0 = 2$. We have $n \leq cn \log n$ for $n \geq n_0$, because n is positive and $1 \leq \log n$ for $n \geq 2$.

- (b) $n^{1/\log(n)} = \Theta(1)$.

SOLUTION: True. Let $n^{1/\log(n)} = x$. Taking log on both sides, we get, $\log(n^{1/\log(n)}) = \log(x)$, $\frac{\log(n)}{\log(n)} = \log(x)$, $1 = \log(x)$, $x = 2$ i.e. $n^{1/\log(n)} = 2$. Therefore, we can choose c_1 as 1 and c_2 as 4 and n_0 as 2 and for $n \geq n_0$, we get $c_1 \leq n^{1/\log(n)} \leq c_2$.

- (c) If

$$f(n) = \begin{cases} 5^n & \text{if } n < 2^{1000} \\ 2^{1000}n^2 & \text{if } n \geq 2^{1000} \end{cases}$$

and $g(n) = \frac{n^2}{2^{1000}}$, then $f(n) = O(g(n))$.

SOLUTION: True. We have $f(n) \leq cg(n)$ when $n \geq 2^{1000}$ and $c = 2^{2000}$. Observe that the exponential growth in f at the beginning doesn't matter, because big- O notation only describes what happens for large values of n .

- (d) For all possible functions $f(n), g(n) \geq 0$, if $f(n) = O(g(n))$, then $2^{f(n)} = O(2^{g(n)})$.

SOLUTION: False. Let $f(n) = 2n$ and $g(n) = n$. Then $f(n) = O(g(n))$ because for $c = 2$, $2n \leq 2n$, $\forall n \geq 0$. However, $2^{f(n)} = O(2^{g(n)})$ is not true because let's assume $\exists c, n_0$ such that $2^{2n} \leq c2^n$ for $\forall n \geq n_0$ implies $2^n \leq c$ which is a contradiction to fact that this statement is true $\forall n \geq n_0$

- (e) $5^{\log \log(n)} = O(\log(n)^2)$

SOLUTION: False. First, note that $5^{\log \log n} = 2^{\log 5 \cdot \log \log n} = (2^{\log \log n})^{\log 5} = (\log n)^{\log 5}$, by properties of exponents and logarithms. So the statement reduces to $\log(n)^{\log 5} = O(\log(n)^2)$. This is false since $\log 5 > 2$.

Specifically, suppose there were values of c and n_0 such that $(\log n)^{\log 5} \leq c(\log n)^2$ for all $n \geq n_0$. Then (assuming without loss of generality that $n_0 > 1$, since if $n_0 \leq 1$ we can just replace n_0 with $\max(n_0, 1)$), we can divide both sides by $(\log n)^2$ and obtain $(\log n)^\epsilon \leq c$ for some positive constant ϵ and all $n \geq n_0$. This means $\log n \leq (\log c)/\epsilon$ for all $n \geq n_0$. However, $\log n$ diverges to infinity as n grows large, producing a contradiction.

(f) $n = \Theta(100^{\log(n)})$

SOLUTION: False. $100^{\log(n)} = n^{\log(100)}$ (because $a^{\log(b)} = b^{\log(a)}$ keeping the log bases same). Therefore, we need to disprove that $n = \Theta(n^{\log(100)})$. We will show that $n = \Omega(n^{\log(100)})$ is false which will be sufficient to prove the claim. Let us say that $\exists c, n_0$ such that $\forall n \geq n_0, n \geq cn^{\log(100)}$, then we get that $n^{\log(100)-1} \leq \frac{1}{c}$, $n \leq \frac{1}{c^{\frac{1}{\log(100)-1}}}$ which is a contradiction to the fact that the statement is true for all $n \geq n_0$.

(g) The following two problems **are not required** but might be fun to think about. We will give you feedback if you do them but they will not affect your grade.

i. (**bonus**) $n|\sin(\pi n/2)| = O(n|\cos(\pi n/2)|)$

SOLUTION: False. Whenever $n \equiv 1 \pmod{2}$, $|\sin(\pi n/2)| = 1$ and $|\cos(\pi n/2)| = 0$. Therefore, $n|\sin(\pi n/2)| > cn|\cos(\pi n/2)|$ for all nonzero values of c , and all odd values of n . Since for any number n_0 , there will always be an odd number greater than n_0 , there cannot exist values of c and n_0 such that $n|\sin(\pi n/2)| \leq cn|\cos(\pi n/2)|$ for all $n \geq n_0$.

ii. (**bonus**) $n|\sin(\pi n/2)| = \Omega(n|\cos(\pi n/2)|)$

SOLUTION: False. Whenever $n \equiv 0 \pmod{2}$, $|\sin(\pi n/2)| = 0$ and $|\cos(\pi n/2)| = 1$. Therefore, $n|\sin(\pi n/2)| < cn|\cos(\pi n/2)|$ for all nonzero values of c , and all even values of n . Since for any number n_0 , there will always be an even number greater than n_0 , there cannot exist values of c and n_0 such that $n|\sin(\pi n/2)| \geq cn|\cos(\pi n/2)|$ for all $n \geq n_0$.

3. **n-naught not needed.** (3 points) Suppose that $T(n) = O(n^d)$, and that $T(n)$ is never equal to ∞ . Prove rigorously that there exists a c so that $0 \leq T(n) \leq c \cdot n^d$ for all $n \geq 1$. That is, the definition of $O(\cdot)$ holds with $n_0 = 1$. [**We are expecting a rigorous proof using the definition of $O(\cdot)$.**]

SOLUTION: Suppose that $T(n) = O(n^d)$. This means that there exists a c and an n_0 so that $T(n) \leq cn^d$ for all $n \geq n_0$.

Let

$$c' = \max \left(c, \frac{T(1)}{1^d}, \frac{T(2)}{2^d}, \dots, \frac{T(n_0 - 1)}{(n_0 - 1)^d} \right)$$

. Then

$$T(n) \leq \begin{cases} T(n) & n < n_0 \\ cn^d & n \geq n_0 \end{cases} \leq \begin{cases} c'n^d & n < n_0 \\ c'n^d & n \geq n_0 \end{cases} \leq c'n^d$$

4. **Fun with recurrences.** (6 points; 1 point each)

Solve the following recurrence relations; i.e. express each one as $T(n) = O(f(n))$ for the tightest possible function $f(n)$, and give a short justification. Be aware that some parts might be slightly more involved than others. Unless otherwise stated, assume $T(1) = 1$. [To see the level of detail expected, we have worked out the first one for you.]

- (z) $T(n) = 6T(n/6) + 1$. We apply the master theorem with $a = b = 6$ and with $d = 0$. We have $a > b^d$, and so the running time is $O(n^{\log_6(6)}) = O(n)$.
- (a) $T(n) = 2T(n/2) + 3n$
- (b) $T(n) = 3T(n/4) + \sqrt{n}$
- (c) $T(n) = 7T(n/2) + \Theta(n^3)$
- (d) $T(n) = 4T(n/2) + n^2 \log n$
- (e) $T(n) = 2T(n/3) + n^c$, where $c \geq 1$ is a constant (that is, it doesn't depend on n).
- (f) $T(n) = 2T(\sqrt{n}) + 1$, where $T(2) = 1$

SOLUTION:

- (a) $T(n) = O(n \log n)$, using the Master Theorem with $a = 2, b = 2, d = 1$.
- (b) $T(n) = O(n^{\log_4 3})$, using the Master Theorem with $a = 3, b = 4, d = 1/2$. We have $a > b^d$, so the answer is $O(n^{\log_b(a)})$.
- (c) $T(n) = O(n^3)$, using the Master Theorem with $a = 7, b = 2, d = 3$. (Notice that the $\Theta(n^3)$ expression is $O(n^3)$ as well, as per the statement in class.) Then $a < b^d$, so the running time is $O(n^d) = O(n^3)$.

(d) $T(n) = O(n^2 \log^2(n))$. To get an idea, we might start working things out:

$$\begin{aligned}
T(n) &= 4T(n/2) + n^2 \log(n) \\
&= 16T(n/4) + n^2 \log(n/2) + n^2 \log(n) \\
&= \dots \\
&= 2^{2k}T(n/2^k) + n^2(\log(n/2^{k-1}) + \dots + \log(n/2) + \log(n)) \text{ where } k = \log_2(n) \\
&= n^2 + n^2 \log\left(\frac{n^k}{2^{k(k-1)/2}}\right) \\
&= n^2 + n^2 \log^2(n) \\
&= O(n^2 \log^2(n)).
\end{aligned}$$

Formally (a formal proof is not required for credit, but here it is), we may proceed by induction. We use as an inductive hypothesis that for all $t > 0$.

$$T(n) = 4^t \cdot T(n/2^t) + n^2(\log(n) + \log(n/2) + \log(n/4) + \dots + \log(n/2^{t-1})).$$

For the base case, notice that when $t = 1$ this is just

$$T(n) = 4 \cdot T(n/2) + n^2 \log(n),$$

which is the provided recurrence relation. For the inductive hypothesis, we observe that

$$T(n) = 4^{t-1} \cdot T(n/2^{t-1}) + n^2(\log(n) + \dots + \log(n/2^{t-2})),$$

and plugging in the recurrence relation we find

$$T(n) = 4^{t-1} \cdot (4T(n/2^t) + (n/2^{t-1})^2 \log(n/2^{t-1})) + n^2(\log(n) + \dots + \log(n/2^{t-2})).$$

Rearranging establishes the inductive hypothesis for the next round. Finally, we plug in $t = \log_2(n)$ to see that

$$T(n) = 4^{\log_2(n)}T(1) + n^2(\log(n) + \dots + \log(2)) = O(n^2 \log^2(n)),$$

as desired.

(e) The Master Theorem implies that

$$T(n) = \begin{cases} O(n^{\log_3 2} \log n) & \text{if } 2 = 3^c \\ O(n^c) & \text{if } 2 < 3^c \\ O(n^{\log_3 2}) & \text{if } 2 > 3^c \end{cases}$$

However, since $c \geq 1$, we always have $2 < 3^c$, so the answer is $O(n^c)$.

(f) To get intuition, we begin by iteratively plugging in the recurrence relation:

$$\begin{aligned}
T(n) &= 2T(\sqrt{n}) + 1 \\
&= 2(2T(n^{1/4}) + 1) + 1 \\
&= 4T(n^{1/4}) + 2 + 1 \\
&= 4(2T(n^{1/8}) + 1) + 2 + 1 \\
&= 8T(n^{1/8}) + 4 + 2 + 1
\end{aligned}$$

Carrying on this way, we see that for $t \geq \log \log(n)$,

$$T(n) = 2^t T(n^{1/2^t}) + \sum_{i=0}^{t-1} 2^i.$$

(To formally prove this, we may do a proof by induction; this is not required for credit for this problem.) Now, for $t = \log \log(n)$, we have $n^{1/2^t} = 2$, and so plugging in $t = \log \log(n)$ we see

$$T(n) = 2^{\log \log(n)} T(2) + \sum_{i=0}^{\log \log(n)-1} 2^{\log \log(n)} = O(\log(n)).$$

5. **Different-sized sub-problems.** (6 points) (Note: this problem will be easier to do after class on Monday 4/16.) Solve the following recurrence relation.

$$T(n) = T(n/2) + T(n/4) + T(n/8) + n,$$

where $T(1) = 1$. [We are expecting a formal proof. You may state your final running time with $O(\cdot)$ notation, but do not use it in your proof.]

SOLUTION: Solution 1: Substitution method. We use the substitution method to prove that $T(n) = O(n)$. Formally, we use the inductive hypothesis that $T(n) \leq 8 \cdot n$ for all $n \geq 1$. To see the base case, we have $T(1) = 1 \leq 8$. Now we proceed by induction. Assuming that $T(k) \leq 8k$ for all $k < n$, we have

$$\begin{aligned} T(n) &= T(n/2) + T(n/4) + T(n/8) + n \\ &\leq 8n/2 + 8n/4 + 8n/8 + n \\ &= 7n + n \\ &= 8n, \end{aligned}$$

which establishes the inductive hypothesis for n . By induction, we conclude that $T(n) \leq 8n$ for all $n \geq 1$. From the definition of $O()$ (with $c = 8$ and $n_0 = 1$) we conclude that $T(n) = O(n)$.

Solution 2: Recursion tree. Level 0 has a cost of n . Level 1 has a cost of $n(\frac{4+2+1}{8}) = \frac{7}{8}n$. Level 2 has a cost of $n(\frac{1}{4} + \frac{2}{8} + \frac{3}{16} + \frac{2}{32} + \frac{1}{64}) = \frac{49}{64}n = (\frac{7}{8})^2 n$. In general, each level has a cost of $7/8$ times the cost of the previous level, because each node on the previous level (with cost x) has children with costs $\frac{x}{2}$, $\frac{x}{4}$, and $\frac{x}{8}$, which sum to $\frac{7}{8}x$.

Accounting for the base case here is tricky, because different branches of the tree are different heights, and the bottom level is not fully populated with nodes. In addition, if you try to find an upper bound by assuming the bottom level is fully populated with nodes, you get a total cost that is greater than the bound we are trying to prove. In order to account for the base case, we may do the following:

- (a) Define $T(x)$ to be 0 when x is less than 1. Then the recurrence $T(n) \leq T(n/2) + T(n/4) + T(n/8) + n$ is valid over all real numbers.
- (b) Produce an infinite “fake recursion tree,” with no base case, and compute the sum of the costs for that tree. In this tree, some of the nodes will have fractional costs, and each node will represent the (possibly fractional) work done by the recursive call. Because of the way the base case is defined, the sum of the costs for this tree is strictly larger than the sum of the costs of the finite recursion tree.

The cost of the fake recursion tree is just $\sum_{i=0}^{\infty} (\frac{7}{8})^i n$, which by the sum of a geometric series, is $O(n)$. Note that this must also be $\Omega(n)$ because the top level alone already costs n .

6. **What’s wrong with this proof?** (8 points) Consider the following recurrence relation:

$$T(n) = T(n - 5) + 10 \cdot n$$

for $n \geq 5$, where $T(0) = T(1) = T(2) = T(3) = T(4) = 1$. Consider the following three arguments.

1. **Claim:** $T(n) = O(n)$. To see this, we will use strong induction. The inductive hypothesis is that $T(k) = O(k)$ for all $5 \leq k < n$. For the base case, we see $T(5) = T(0) + 10 \cdot 5 = 51 = O(1)$. For the inductive step, assume that the inductive hypothesis holds for all $k < n$. Then

$$T(n) = T(n - 5) + 10n,$$

and by induction $T(n - 5) = O(n - 5)$, so

$$T(n) = O(n - 5) + 10n = O(n).$$

This establishes the inductive hypothesis for n . Finally, we conclude that $T(n) = O(n)$ for all n .

2. **Claim:** $T(n) = O(n)$. To see this, we will use the Master Method. We have $T(n) = a \cdot T(n/b) + O(n^d)$, for $a = d = 1$ and

$$b = \frac{1}{1 - 5/n}.$$

Then we have that $a < b^d$ (since $1 < 1/(1 - 5/n)$ for all $n > 0$), and the master theorem says that this takes time $O(n^d) = O(n)$.

3. **Claim:** $T(n) = O(n^2)$. Imagine the recursion tree for this problem. (Notice that it’s not really a “tree,” since the degree is 1). At the top level we have a single problem of size n . At the second level we have a single problem of size $n - 5$. At the t ’th level we have a single problem of size $n - 5t$, and this continues for at most $t = \lfloor n/5 \rfloor + 1$ levels. At the t ’th level for $t \leq \lfloor n/5 \rfloor$, the amount of work done is $10(n - 5t)$. At the last level the amount of work is at most 1. Thus the total amount of work done is at most

$$1 + \sum_{t=0}^{\lfloor n/5 \rfloor} 10(n - 5t) = O(n^2).$$

- (a) (3 points) Which, if any, of these arguments are correct? **[We are expecting a single sentence stating which are correct.]**

SOLUTION: The first two are incorrect. The third is correct.

- (b) (5 points) For each argument that you said was incorrect, explain why it is incorrect. If you said that all three were incorrect, then give a correct argument. **[We are expecting a few sentences of detailed reasoning for each incorrect algorithm; and if you give your own proof we are expecting something with the level of detail of the proofs above—except it should be correct!]**

SOLUTION: (Two points for why argument 1 is incorrect, one point for argument 2). The first argument is incorrect because it mis-uses big-oh notation. The inductive hypothesis doesn't even make sense: unpacking the $O(\cdot)$ notation, it would say: "For all $k < n$, there exists a c and n_0 so that for all $k \geq n_0$, $T(m) \leq ck$." That is, the definition of $O(\cdot)$ requires something to hold for arbitrarily large n , but the hypothesis requires a bounded n .

Another way to look at this is to say that the constant c that's implicitly hiding inside the $O(\cdot)$ notation might depend on n , because each time we establish the inductive hypothesis, we re-establish the $O(\cdot)$ notation. But c isn't allowed to depend on n .

The second argument is incorrect because it mis-uses the Master Theorem. In the proof of the Master Theorem that we saw in class, it was important that a, b, d be constants, independent of n . But in this proof, we have chosen b to depend on n .